View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Digital library of Brno University of Technology

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ TVORBA SLOVNÍKŮ Z PŘEKLADOVÝCH TEXTŮ

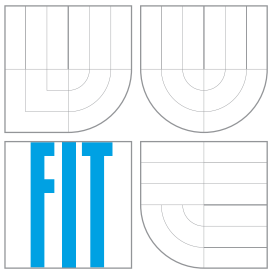
BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

LENKA SUMBALOVÁ

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

AUTOMATICKÁ TVORBA SLOVNÍKŮ Z PŘEKLADOVÝCH TEXTŮ

AUTOMATIC CREATION OF DICTIONARIES FROM TRANSLATIONS

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

LENKA SUMBALOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR

doc. RNDr. PAVEL SMRŽ, Ph.D.

BRNO 2013

Abstrakt

Cílem této bakalářské práce bylo vytvořit systém pro automatickou tvorbu slovníků z překladových testů. Je popsána implementace systému, který generuje česko-anglický slovník ze zarovnaného paralelního korpusu a shrnut výsledek práce. Dále je analyzován paralelní korpus CzEng, který byl použit jako zdroj dat pro slovníky a vysvětleny teoretické pojmy související s touto problematikou.

Abstract

The aim of this bachelor thesis was to make a system for automatic creation of dictionaries from translations. It describes the implementation of a system that generates Czech-English dictionary from the aligned parallel corpus and summarizes the results. It also analyzed CzEng parallel corpus, which was used as the data source for dictionaries and explainS the theoretical concepts related to this topic.

Klíčová slova

slovník, korpus, zarovnání, strojový překlad, lemmatizace

Keywords

dictionary, corpus, alignment, machine translation, lemmatization

Citace

Lenka Sumbalová: Automatická tvorba slovníků
z překladových textů, bakalářská práce, Brno, FIT VUT v Brně, 2013

Automatická tvorba slovníků z překladových textů

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením pana doc. RNDr. Pavla Smrže, Ph. D. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

.....
Lenka Sumbalová
15. května 2013

Poděkování

Chtěla bych poděkovat doc. RNDr. Pavlu Smržovi, Ph.D. za odborné vedení práce.

© Lenka Sumbalová, 2013.
Tato práce vznikla jako školní dílo na Vysokém učení technickém v Brně, Fakultě informačních technologií. Práce je chráněna autorským zákonem a její užití bez udělení oprávnění autorem je nezákonné, s výjimkou zákonem definovaných případů.

Obsah

Kapitola 1

Úvod

S pokrokem v oblasti komunikačních technologií a dopravy se stále zmenšují vzdálenosti mezi lidmi z různých zemí. Lidé cetují do zahraničí, komunikují spolu prostřednictvím internetu. Aby se spolu mohli dorozumnět lidé ze zemí, kde se mluví rozdílným jazykem, je třeba se cizím jazykům učit. Jednou z nejdůležitějších pomůcek pro naučení cizího jayzyka bývá dvojjazyčný slovník.

S potřebou komunikace mezi lidmi mluvicími různými jazyky také souvisí vývoj oblasti zpracování přirozeného jazyka, konkrétně strojového překladu. Základní zdroj dat pro práci překladačů je opět slovník.

Manuální tvorba slovníků je velmi náročná, slovník může vznikat i několik let. Je náročné obsáhnout veškerou slovní zásobu jazyka, také je třeba brát v potaz neustálý vývoj jazyků a slovníky aktualizovat. Tuto práci mohou ulehčit nástroje pro automatizované vytváření slovníků. Slovník může být vytvářen např. z paralelních textů, textů ve více jazycích vyjadřujících tutéž skutečnost. Pokud jsou použity vhodné texty, vzniklý slovník bude obsahovat aktuální slovní zásobu daných jazyků

Tato práce se věnuje vytváření slovníků ze zarovnaných paralelních textů, konkrétně tvorbě anglicko-českého slovníku. Jako zdroj paralelních textů byl použit česko-anglický zarovnaný paralelní korpus CzEng vytvořený na Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty University Karlovy.

Ve druhé kapitole této práci jsou popsány pojmy související s touto problematikou, strojový překlad, paralelní texty a jejich zarovnávání a lemmatizace. Třetí kapitola popisuje výše zmíněný paralelní korpus CzEng. Čtvrtá kapitola popisuje postup tvorby systému pro vytváření slovníku z paralelních textů, kapitola pátá je zaměřena na shrnutí výsledků a srovnání vytvořených slovníků se slovníkem existujícím.

Kapitola 2

Rozbor tématu

2.1 Strojový překlad

Strojovým překladem rozumíme automatizovaný překlad mezi přirozenými jazyky, který provádí počítač. Oblast strojového překladu se stále vyvíjí, existuje mnoho systémů, jejichž výstupy však nejsou dokonalé. Jsou však dostatečně kvalitní jako pomoc pro lidské překladatele, pro urychlení jejich práce. Dále mohou sloužit k základní orientaci pro lidi, kteří daný cizí jazyk neznají. Strojový překlad není možno použít k překladu literáních textů. Je však možno překládat např. technické manuály, vědecké dokumenty a komerční prospekty. Největší úspěch mají překladače v případě překladu odborných textů z úzce omezeného oboru; překlad je také jednodušší, pokud nejsou mezi jazyky velké rozdíly (např. pokud se jedná o dva slovanské jazyky).

Kvalita překladu může být zlepšena několika způsoby. Překladový systém může být vytvářen pro překlady týkající se omezené oblasti nebo jen pro určitý typ dokumentů. Další způsob je omezení vstupních textů, např. o zákaz výskytu homonym nebo použití jen omezené slovní zásoby. Dále je možné texty předem označit, přiřadit ke slovům slovní druhy či gramatické kategorie, přiřadit slovům jejich základní tvary atd.

Překladové systémy mohou být jak dvojjazyčné, tak vícejazyčné. Může se jednat o překlad jen v jednou směru (např. pouze z češtiny do angličtiny), nebo o obousměrný překlad.

Můžeme se setkat se čtyřmi základními strategiemi strojových překladačů. Historicky nejstarší přístup je přímý překlad. Tyto systémy jsou nutně dvojjazyčné a jednosměrné; jedná o se o překlad z konkrétního zdrojového do rovněž pevně daného cílového jazyka. Systém většinou obsahuje dvojjazyčný slovník a program pro analýzu zdrojového jazyka, ten je analyzovám jen do té míry, nakolik je to potřeba pro překlad do jazyka cílového. Druhou možností je využití tzv. „mezijazyka“ – nejdřív je text ze zdrojového jazyka přeložen do tohoto „mezijazyka“ a z něj je přeložen do cílového jazyka, tyto dvě fáze překladu jsou na sobě nezávislé. Výhodou tohoto přístupu je možnost použít jej v systému pro překlad mezi mnoha jazyky – není potřeba mít programy pro překlad mezi každými dvěmi jazyky. Jako mezičlánek překladu se může použít uměle vytvořený jazyk jako Esperanto nebo „universální“ jazykově nezávislá slovní zásoba. Další přístup je překlad skládající se ze tří fází. Pro zdrojový i cílový jazyk je použita jeho zjednodušená verze, nejdřív se překládá do zjednodušeného zdrojového jazyka, z něj do zjednodušeného cílového jazyka a konečně do cílového jazyka. Jsou obvykle použity tři typy slovníků a větší množsví gramatik. Poslední přístup se nazývá empirický. Tento přístup využívá velké množství dat v obou jazycích; může se jednat o paralelní korpus nebo příklady překladů [5].

Dále můžeme dělit překladové metody na metody řízené pravidly, metody založené na

příkladech, statistické metody a jejich kombinace. Tyto různé metody zde nyní popíšu.

2.1.1 Metody řízené pravidly

Tato překladová metoda je nejstarší. Využívá lingvistické informace o zdrojovém i cílovém jazyce, je založena na morfologické, syntaktické a sémantické analýze obou jazyků. Systém pro překlad se skládá ze slovníků, gramatických pravidel a programu, který tato pravidla aplikuje. Tento přístup využívají např. systémy Systran, Eurotra, Apertium.

Věta ve zdrojovém jazyce je reprezentována stromovou strukturou vytvořenou na základě gramatických pravidel. Pomocí slovníku se slova zdrojového jazyka přeloží do cílového a opět se vytvoří strom, tentokrát aplikací pravidel cílového jazyka.

Výhodou tohoto přístupu je hloubková syntaktická a sémantická analýza. Nevýhoda je náročnost na lingvistické znalosti a potřeba velkého množství pravidel pro pokrytí veškerých aspektů obou jazyků.[18]

2.1.2 Metody založené na příkladech

Přístup k překladu založený na příkladech využívá existující překlady mezi zdrojovým a cílovým jazykem (může se jednat o paralelní korpus). Tyto překlady jsou vytvořené člověkem, počítač podle nich může vytvářet překlady další.

V první fázi se pro vstupní větu ve zdrojovém jazyce musí nalézt odpovídající věta v paralelním korpusu – věta, která je té vsupní co nejvíce podobná, shoduje se v co nejvíce slovech. Následuje zarovnání, na jehož základě se určí, které části vět z korpusu je možno použít pro překlad. Poté se z takových částí poskládá výstupní věta.[19]

2.1.3 Statistické metody

Překlady statistickou metodou jsou generovány na základě statistických modelů, jejichž parametry jsou získané pomocí analýzy dvojjazyčného korpusu. Jedná se o nejčastěji používané metody, na tomto principu pracuje např. Google překladač.

Statistické metody předpokládají, že každá věta cílového jazyka je překladem věty ze zdrojového jazyka s jistou pravděpodobnstí; překlad s nejvyšší pravděpodobností je nejlepší. Hlavními úkoly statistického překladu je tedy odhad pravděpodobnosti a nalezení překladu s nejvyšší pravděpodobností.

Statistické překladače mohou být založeny na slovech, frázích nebo syntaxi. Systémy založené na slovech dnes nejsou příliš používány. Paralelní texty jsou zarovnány na slova a podle četností různých překladů určitého slova ve zdrojovém jazyce do cílového se určují pravděpodobnosti těchto překladů. Častější je přístup založený na frázích, místo samotných slov se používají jejich sekvence. Překlad založený na syntaxi se zaměřuje na překlad syntaktických jednotek.

2.1.4 Hybridní metody

Hybridní překladače využívají zároveň pravidla i statistiku. Může se jednat o překlad založený na pravidlech, který je poté zkorigován pomocí statistiky. Jiným přístupem je předzpracování textu pomocí pravidel, překlad statistickým překladačem a případně další úpravy opět pomocí pravidel.

2.2 Paralelní texty a jejich zarovnávání

Paralelní texty jsou stejné texty ve více jazycích, většinou se jedná o originální text a jeho překlad (překlady). Zarovnáváním paralelních textů rozumíme identifikaci vět a slov v obou (všech) jazykových verzích textu, které jsou významově ekvivalentní.

Známým paralelním textem je Rosetská deska. Jedná se o desku pocházející ze starověkého Egypta, která obsahuje text v egyptštině a řečtině; na jejím základě Jean-Francois Champollion rozluštil hieroglyfy. Podobných příkladů, které napomohly k rozluštění starověkých jazyků, najdeme v historii více. Další paralelní texty, které nám historie nabízí jsou např. různé smlouvy, náboženské texty, literatura, atd. První pokusy použít paralelní texty pro strojový překlad se vyskytly v padesátých letech dvacátého století, tehdejší možnosti však byly omezené. Dále se obor rozvíjel v osmdesátých letech; výzkumná centra shromažďovala paralelní texty, vznikly první metody jejich automatického zarovnávání.[12]

Zarovnané paralelní texty mají mnoho využití. Jedná se např. o vytváření slovníků a dvojjazyčných seznamů terminologie, strojový překlad, hledání příkladů pro počítačové učení.

Kolekce obsahující mnoho paralelních textů se nazývají paralelní korpusy. Korpusy mohou obsahovat texty ve dvou jazycích (jazyk originálních textů a jejich překlad), taktéž může existovat paralelní korpus obsahující texty ve více jazycích. Směr překladu nemusí být pro všechny texty stejný, např. v případě česko-anglického paralelního korpusu může korpus obsahovat jak anglické texty a jejich české překlady, tak i naopak české texty a jejich anglické překlady.

2.2.1 Zarovnání paralelních textů

Zarovnávání paralelních textů má mnoho využití. Je nutné pro statistický strojový překlad, automatickou extrakci překladových ekvivalentů z textu, automatickou tvorbu slovníků z těchto textů. Jedná se o nalezení korespondence mezi oběma stranami textu, ať už na úrovni slov, frází, odstavců nebo vět. Pro zarovnání na nižší úrovni (na slova) je třeba nejdříve texty zarovnat na úrovni vyšší (na věty).

Bez ohledu na to, zda se jedná o úroveň vět či slov rozlišujeme několik typů zarovnání. Texty mohou být zarovnány `1:1`, znamená to, že každé větě (každému slovu) zdrojového textu odpovídá právě jedna věta (jedno slovo) v textu cílovém. Další typy zarovnání jsou `1:0`, kdy jsou v cílovém textu některé věty (slova) ze zdrojového vynechány, `0:1`, kdy je naopak ve zdrojovém textu věta (slovo), která v cílovém textu nemá svůj ekvivalent. Nejobecnější případ je zarovnání `m:n`.

Proces zarovnání

Zarovnání paralelních textů se skládá z několika fází. Nejprve je nutné shromáždit korpus, následuje zpracování vstupu, zarování na dokumenty, věty, slova.

Zdrojů pro vytváření zarovnaného korpusu je možno nalézt mnoho. Může se jednat o literární texty (knihy publikované elektronicky; originál a překlad), náboženské texty (např. Bible je přeložena do více než 400 jazyků), mezinárodní právo (např. dokumenty Evropské unie, Všeobecná deklarace práv a svobod, Kjótský protokol), titulky k filmům, dokumentace k softwaru (dostupná ve více jazycích), dvojjazyčné časopisy (např. National Geographic), webové stránky (s možností vybrat si jazyk zobrazení).

Získané zdrojové texty jsou v různých formátech, záleží na zdroji. Mohou se vyskytovat ve formátu `pdf`, `html` (webové stránky), `SubRip` (titulky), prostého textu, atd. Texty v `pdf`

formátu jsou obvykle převáděny na prostý text, může být neformátovaný, případně `html` nebo `xml`. Bohužel převod není zcela dokonalý, často zůstávají čísla stránek či záhlaví. Texty ve formátu `html` či `xml` jsou pro zpracování výhodnější. Navíc lze v případě textů ve formátu `html` využít některé `html` značky pro zarovnání (např. nadpis, zvýrazněný text).

První fází samotného zarovnání je zarovnání na úrovni dokumentů. Hodně dokumentů neobsahuje žáné informace o tom, které z nich spolu korespondují. Tato situace je typická pro automaticky získaný text z webových stránek. Často však jednotivé jazykové verze těchto stránek obsahují v názvu zkratku jazyka, jindy je možné zarovnat dokumenty na základě metainformací nebo cestě k jejich původnímu uložení.

Následuje zarovnání sekcí, odstavců nebo nejčastěji vět. Pro větné zarovnání je třeba texty nejdřív na věty rozdělit. Nejjednodušší přístup je dělení na věty na základě interpunkce (tečka, vykřičník, otazník). Toto však často nestačí, je třeba brát v potaz jiné zakončení věty nebo možné zkratky ukončené tečkou, to je závislé na konkrétním jazyce.

Vstupem větného zarovnání jsou texty určené k zarovnání, někdy také slovník. Většinou zarovnání začíná výpočtem jeho skóre, snahou najít záchytné body pro zahájení zarovnání. Skóre může být vypočteno na základě délky slov, slovníku, syntaxe, podobnosti termínů. Po nalezení záchytných bodů se proces opakuje, dokud je možné najít nové větné páry.

Před zarovnáním na slova je třeba věty na slova rozdělit. Toho jde dosáhnou například pomocí rozdělení znaků na slovní a neslovní. V některých jazycích, jako je například čínština, je však situace komplikovanější.

Zarovnání na slova je podobné větnému zarovnání, je však komplikovanější. Slovosled se často hodně liší, slova zarovnaná na sebe nemusí mít stejné slovní druhy, je třeba se vypořádat se slovními spojeními a frázemi. [11]

Zarovnání na věty

Existují různé algoritmy a nástroje pro větné zarovnání. Způsoby zarovnání můžeme dělit na založené délce, slovníku nebo částečné podobnosti.

První systém pro zarovnání vycházel z předpokladu, že pokud jsou na sebe věty zarovnané, jsou na sebe zarovnaná i slova v nich. Počítal s pravděpodobnostmi zarovnání vět na sebe; je pravděpodobné, že první a poslední věty v textech patří k sobě, ostatní věty mají pravěpodobnost nižší. Slova jsou klasifikována: slova se stejnou klasifikací jsou spárována a považována za záchytné body spárovaných vět. Algoritmus běží v iteracích, dokud není zarovnání nalezeno.

Popsaný algoritmus není však vhodný pro objemné korpusy, proto vznikly další přístupy založené na délce vět. Tyto přístupy se liší ve způsobu měření délky vět – může se jednat o počet slov nebo počet znaků. Bylo zjištěno, že existuje relativně fixní poměr mezi délkami vět v jakýchkoliv dvou jazycích.

Další metody jsou založeny na slovníkových informacích. Překládají se některá slova ve větách a na jejich zakladě jsou vybrány věty, které k sobě patří.

Jiné algoritmy využívají princip podobnosti textů: hledají se tokeny, které jsou graficky či jinak podobné; může se jednat o čísla, data, vlastní jména, interpunkční znaménka atd. [11]

Zarovnání na slova

Nejběžnější přístup k zarovnání na slova je založen na generativních modelech. Tyto modely popisují proces vytvoření věty v jednom jazyce z věty v jazyce druhém, pomocí tohoto modelu je možné rekonstruovat původní větu. Generativní modely mohou využívat metody

učení bez učitele, není potřeba mít k dispozici ručně zarovnané texty, naopak potřebujeme velké množství vět k zarovnání. Nevýhodou tohoto přístupu je složitost modelů. Modely nejsou příliš flexibilní a přidání nových vlastností je problematické.

Druhý přístup je diskriminační. V porovnání s předchozím je mnohem flexibilnější. Požaduje však ručně zarovnané věty pro učení s učitelem.

Dalším přístupem jsou heuristické metody, které využívají podobnostní funkce. Ve srovnání s výše uvedenými přístupy jsou tyto metody mnohem jednodušší, jejich výkon je však horší.[3]

Víceslovné fráze

Věty nikdy nejsou zarovnané na slova `1:1`. Často je slovo v jednom jazyce reprezentováno ve druhém jazyce víceslovnou frází, případně se jedná o překlad mezi dvěma frázemi.

Většina zarovnávacích metod považuje větu za sekvenci slov a korespondence mezi frázemi generuje s využitím heuristik. Některé metody hledají zarovnání na fráze až po slovním zarovnání, to samotné však není bezchybné, proto tento přístup nepřináší nejlepší výsledky. Některé modely využívají strukturální informace od začátku celého procesu zarovnání. Strukturální zarovnávací metody využívají heuristická pravidla pro vyřešení nejednoznačností při zarovnání. Dále existují pravděpodobnostní metody založené na stromové struktuře vět. Tyto metody přeuspořádávají, vkládají nebo odstraňují podstromy frází v jednom jazyce, aby to bylo reprodukovatelné na straně druhé. Jiný model se zaměřuje na závislostní vztahy.[9]

2.2.2 Nástroje pro zarovnání

NATools

Jedná se o množinu nástrojů pro práci s paralelními texty, tedy i jejich zarovnání. Obsahuje nástroj pro větné i slovní zarovnání. Nástroje byly vyvinuty na Universitě v Minho.[11]

Hunalign

Nástroj hunalign slouží k zarovnání paralelních textů na úrovni vět. Pokud dostane hunalign na vstup slovník, využije jej společně s délkou vět. Jinak nejdřív využívá pouze délku vět a automaticky slovník vytvoří na základě tohoto zarovnání, potom provede druhé zarovnání s použitím slovníku. Není schopen se vypořádat se změnou v pořadí vět. Je napsán v jazyce C++ a může být spuštěn v jakémkoliv operačním systému.[26]

Giza++

Jedná se o rozšíření staršího nástroje Giza. Používá se pro zarovnání na slova, využívá statistické zarovnávací metody.[11]

WinAlign

WinAlign je komerční nástroj vyvinutý pro profesionální překladače. Využívá předchozí překlady jako překladovou paměť pro další zarovnání.[11]

2.2.3 Existující paralelní korpusy

Pro generování slovníků jsem využila paralelní korpus CzEng, jehož popisu se budu věnovat v následující kapitole. Zde se zaměřím na další existující paralelní korpusy.

2.2.4 JRC-Acquis

JRC-Acquis je vícejazyčný paralelní korpus obsahující legislativu evropské unie, která je přeložená do 23 jazyků: bulharštiny, češtiny, dánštiny, němčiny, řečtiny, angličiny, španělštiny, estonštiny, finštiny, francouzštiny, maďarštiny, italštiny, litevštiny, lotišštiny, maltánštiny, holandštiny, portugalštiny, rumunštiny, slovenštiny, slovinštiny a švédštiny; korpus je tedy jazykově hodně bohatý. Do korpusu jsou texty přidávány pokud jsou dostupné v alespoň deseti z těchto jazyků. Jedná se o největší paralelní korpus, jak z pohledu velikosti (obsahuje 4 350 447 zarovnaných dokumentů), tak podle počtu jazyků. Jeho velkou výhodou je velký počet neobvyklých jazykových kombinací (maltánština-estonština, slovinština-finština, atd.).[25]

2.2.5 Europarl

Tento paralelní korpus je získán ze sborníku Evropského parlamentu. Stejně jako předchozí je hodně jazykově bohatý, obsahuje texty ve 21 jazycích zemí Evropské unie. Cílem zpracování textů bylo vytvoření paralelního textu zarovnaného na věty pro statistické strojové překladové systémy.[29]

2.2.6 Opus

Opus je otevřený paralelní korpus, kolekce přeložených textů z webu. Autoři korpusu se snažili získat a zarovnat data dostupná zdarma na internetu, přidat do textu lingvistické anotace a vytvořit tak volně dostupný paralelní korpus. Korpus se stále rozrůstá, přibývají data za dalších zdrojů.[22]

2.2.7 InterCorp

Korpus InterCorp je výstupem projektu Filosofické fakulty University Karlovy s cílem vybudovat paralelní synchronní korpus, který pokrývá co nejvíce jazyků. Objem textů, obsah anotace i počet zařazených jazyků se s každou verzí zvyšuje. Jeho jádrem jsou ručně zarovnané literární texty, obsahuje také automaticky zpracované texty. Jako tzv. pivot zde slouží čeština, každý text zde má svou českou verzi. [24]

2.2.8 Další paralelní korpusy

Mezi další korpusy patří např. Glosbe (vícejazyčný paralelní korpus s vyhledávacím rozhraním), Parasol (paralelní korpus vytvořený z beletrie ve slovanských a několika dalších jazycích), Nunavut Hansard (anglicko-inuitský paralelní korpus), TradooIT (anglicko-francouzko-španělský korpus), TERMSEARCH (anglicko-rusko-francouzský paralelní korpus).

2.3 Lemmatizace

Lemmatizace je proces hledání lemmatu zadaného slova. Lemma je základní tvar slova, který se vyskytuje např. ve slovnících (např. lemma slova `kočce` je `kočka`, lemma slova `maluje` je `malovat`). Využívá se např. při dolování znalostí, klasifikaci textu, vyhledávání v textu, fultextovém vyhledávání na internetu. Je vhodné ji použít i při vytváření slovníku z paralelního textu, protože do slovníku patří lemma, nikoliv libovolný jiný tvar slova. Lemmatizace je jedna ze základních metod zpracování přirozeného jazyka.

V češtině je lemma v případě slovních druhů, které se skloňují, první pád jednotného čísla (u přídavných jmen první stupeň, mužský rod), v případě sloves je to infinitiv, lemma příslovcí je jejich první stupeň. U dalších slovních druhů je kvůli jejich neohebnosti každé slovo lemmatem. V případě angličtiny je to podobné, situace je však mnohem jednosduší, protože slova nemají tolik různých tvarů jako v češtině (např. u podstatných jmen se rozlišuje jen jednotné a množné číslo).

Pojem lemmatizace bývá často zaměňován s pojmem stemming. Tyto procesy jsou však rozdílné. Zatím co lemmatizace hledá ke slovům lemmata (normalizované tvary), stemming hledá kmen slova. Lemma a kmen jsou rozdílné pojmy: zatímco lemma je normalizovaný tvar slova, kmen je tvar, který zůstane po odtržení všech flektivních morfémů; např. kmen slova `cvičením` je `cvičen`, zatímco jeho lemma je `cvičení`.

Přístupy k lemmatizaci jsou různé. Mohou být založeny na morfologické analýze či strojovém učení.

Některé algoritmy pracují na základě využití hrubé síly, prohledávají slovník reprezentovaný tabulkou, který obsahuje mapování základních tvarů a skloňovaných tvarů slov. Nevýhodou je velká paměťová náročnost, navíc je kvůli neustálému vývoji jazyka nemožné, aby slovník obsahoval všechny možné tvary všech slov. Proto se tyto algoritmy často používají v kombinaci s jinými.

Jiné algoritmy zpracovávají přípony slov podle určitých pravidel (na základě morfologických pravidel daného jazyka). Ani tyto algoritmy se však nedají vždy použít, v případě nepravidelného tvoření slov nefungují. Většina lemmatizačních algoritmů je založena na určení slovního druhu a následné aplikaci pravidel, která jsou pro každý slovní druh jiná.

Stochistické algoritmy využívají pravěpodobnosti, patří k nim algoritmy strojového učení, které na základě tabulky slov a jejich lemmat trénují lemmatizaci a použijí vždy tvar s nejvyší pravděpodobností výskytu.

Uvedené algoritmy se často kombinují, vznikají tak hybridní algoritmy, např. můžou používat algoritmus zpracování přípon pro většinu slov a tabulku pro nepravidelná.[6]

2.4 Tvorba slovníků

Překladové slovníky jsou základní a velmi důležitá báze znalostí pro konstrukci automatických překladových systémů. Automatická tvorba slovníků může využívat různé zdroje, např. paralelní korpusy nebo paralelní webové stránky.

Metody generování strojově překladových slovníků z paralelních textů používají statistické i lingvistické informace. Statistické metody jsou užitečné pro zarovnání textů na věty a slova. Tyto metody je výhodné použít pro velké množství textů a s velkými frekvencemi výskytů slov. Také lingvistické metody mohou být použity pro nalezení korespondujících slov nebo frází. Jako zdroj lingvistických informací mohou používat např. existující překladový systém. Např. pro složené fráze je možné hledat jejich protějšky pomocí překladu slov

existujícím překladačem a následným hledáním největší podobnosti. Statistické a lingvistické metody také mohou být kombinovány.[7]

Kapitola 3

Paralelní korpus CzEng a jeho analýza

Jako vstupní data jsem v projektu používala paralelní korpus CzEng, konkrétně verzi 1.0. Jedná se o zarovnaný (jak slova, tak věty) česko-anglický paralelní korpus vytvořený na Ústavu formální a aplikované lingvistiky Matematicko-fyzikální fakulty Karlovy university, který je volně dostupný pro nekomerční účely. Při práci s korpusem jsem využila článek Ondřeje Bojara a Zdenka Žabokrtského [2].

Texty v korpusu pocházejí z různých zdrojů. Tyto zdroje je možné rozdělit do osmi kategorií: beletrie, legislativa Evropské unie, titulky k filmům, paralelní webové stránky, technická dokumentace, zprávy, projekt Navajo. Encyklopedie Navajo je projekt využívající překladač webových stránek, obsahuje strojové překlady článků z anglické wikipedie do češtiny [20].

Paralelní korpus CzEng obsahuje 15 136 126 vět. Počet vět v jednotlivých sekcích je uveden v tabulce (3.1). Kromě toho jsou v tabulce uvedeny počty slovních jednotek (slova, interpukce, symboly, atd.). V případě analytické roviny (a-layer) se jedná o všechny slovní jednotky, v tektogramatické rovině (t-layer) jsou obsažena jen významová slova [2].

		Čeština		**Angličtina**	
Kategorie zdrojů	**Věty**	**a-layer**	**t-layer**	**a-layer**	**t-layer**
Beletrie	4 335 183	57 176 714	41 142 003	64 264 229	38 690 193
Legislativa EU	3 992 551	78 022 137	56 446 093	87 488 870	52 717 871
Titulky k filmům	3 076 887	19 571 940	14 614 899	23 353 842	14 917 777
Paralelní webové stránky	1 883 804	30 891 696	23 140 879	35 454 657	22 057 255
Technická dokumentace	1 613 297	16 015 336	11 941 650	16 836 098	11 207 157
Zprávy	201 103	4 280 039	3 207 858	4 736 751	2 963 451
Projekt Navajo	33 301	484 453	363 317	556 702	343 649
Celkem	15 136 126	206 442 315	150 856 699	232 691 149	142 897 353

Tabulka 3.1: Velikost dat v paralelním korpusu CzEng 1.0

3.1 Struktura korpusu

Korpus CzeEng 1.0 je rozdělen do bloků – sekvencí nejvýše patnácti po sobě jdoucích vět pocházejících z jednoho zdroje. Český autorský zákon dovoluje využití krátkých citací publikované práce pro výukové či výzkumné důvody. Z tohoto důvodu, aby bylo zabráněno rekonstrukci původního textu, není použit veškerý text zdojů. Dokumenty byly rozděleny na krátké části, zpřeházeny a byly odstraněny všechny informace, které by napomáhaly určit původní pořadí. Kromě toho nemohou být původní texty rekonstruovány také proto, že se berou v potaz jen věty zarovnané 1-1, tudíž dochází ke ztrátám dat. Další ztráty dat způsobuje filtrování větných párů.

Bloky jsou zpřeházeny, očíslovány a rozděleny do očíslovaných souborů. Každý soubor obsahuje 200 větných párů. Tyto soubory jsou rozděleny do 100 stejně velkých sekcí pojmenovaných `00train` – `97train`, `98dtest`, `99etest`. Poslední dvě sekce slouží k vývojářským a testovacím účelům, ke generování slovníků je nepoužívám. [2].

3.2 Roviny reprezentace vět

Data v korpusu jsou reprezentována na dvou úrovních – analytické a tektogramatické. Tyto roviny nyní popíšu [28].

3.2.1 Analytická rovina

Věty jsou na této rovině reprezentovány orientovaným stromem, každému uzlu stromu odpovídá jedna slovní jednotka (slovo, číslo nebo interpunkce). Hrany mezi uzly vyjadřují vztahy mezi těmito slovními jednotkami. Může se jednat o závislostní vztah či jiné lingvistické jevy. Navíc jsou uzly lineárně uspořádány tak, aby bylo zachováno pořadí slovních jednotek ve větě [28].

3.2.2 Tektogramatická rovina

V tektogramatické rovině je opět věta reprezentována stromem, tentokrát však strom zachycuje hloubkovou strukturu věty. Uzly již netvoří všechny slovní jednotky, ale pouze významová slova. Naopak jsou zde uzly, které žádnému slovu neodpovídají (např. nevyjádřený podmět). Je orientována více sémanticky než rovina analytická. Úkolem této roviny je zdůrazňovat podobnosti mezi jazyky, věty ve dvou různých jazycích reprezentované tektogramatickou rovinou jsou si podobnější, než kdyby byly reprezentovány analyticky.

S touto rovinou také souvisí pojem t-lemma. Jedná se o lemma na tektogramatické rovině, které se od morfologického lemmatu liší. Každému uzlu tektogramatického stromu je přiřazen funktor, který popisuje syntakticko-sémantický vztah uzlu k jeho rodiči. Dalším pojmem je formém, technická zkratka zjednodušující prohledávání kropusu; popisuje, jak je uzel reprezentován v povrchové podobě věty [4].

3.3 Formáty

Paralelní korpus CzEng 1.0 je dostupný ve třech formátech – treex formát, export formát, plaintext formát. V následujících podkapitolách tyto formáty popíši, zaměřím se zejména na export formát, jelikož můj systém pro generování slovníků pracuje s tímto formátem paralelního textu. [2]

3.3.1 Treex formát

Treex formát je primární formát korpusu, obsahuje největší množství informací. Jedná se o text ve formátu XML, který je možné zpracovávat s využitím treex platformy.

3.3.2 Export formát

Tento formát není založený na XML jako předchozí, ale stále obsahuje dostatečné množství informací. Jedná se o řádkově orientovaný formát, každému větnému páru odpovídá jeden řádek. Tyto řádky jsou pomocí tabulátorů rozděleny do sedmnácti sloupců. Následuje popis těchto sloupců společně s příklady:

- ID větného páru – skládá se z označení oblasti, čísla bloku, názvu train/dev/test sekce, čísla souboru, čísla věty v rámci souboru.

  ```
  subtitles-b2-00train-f00001-s8
  ```

- Skóre filtrování – indikuje kvalitu větného páru. Perfektní větný pár má skóre 1, páry se skóre menším než 0,3 nejsou do korpusu zařazeny.

- Česká věta

  - Česká věta z pohledu analytické roviny – každé slovo věty se skládá z několika částí oddělených „|“: slovo ve tvaru vyskytujícím se v textu, lemma, morfologická značka, index slova ve větě, index předchůdce v analytickém stromě, syntaktická funkce. Např.:

    ```
    Zachránil|zachránit_:W|VpYS---XR-AA---|1|0|Pred mi|já|PH-S3--
    1-------|2|1|Obj můj|můj_(přivlast.)|PSYS1-S1-------|3|5|Atr
    milovaný|milovaný_(*2t)|AAIS1----1A----|4|5|Atr krk|krk|NNIS1
    -----A----|5|1|Obj .|.|Z:----...
    ```

  - Česká věta z pohledu tektogramatické rovniny – opět se každé slovo skládá z několika částí oddělelných „|“: t-lemma, funktor, index ve stromě, index předchůdce, typ uzlu, formém a další atributy. Např.:

    ```
    zachránit|PRED|1|0|complex|v:fin|v|-|neg0|ant|ind|decl|-|cpl|
    -|-|disp0|-|it0|-|-|res0|-|-|1|-|- #PersPron|ADDR|2|1|complex|
    n:3|n.pron.def.pers|sg|-|-|-|-|-|-|-|-|nr|-|1|basic|-|-|-...
    ```

  - Korespondence mezi českou analytickou a tektogramatickou rovinou pro významová slova.

  - Korespondence mezi českou analytickou a tektogramatickou rovinou pro pomocná slova.

- Anglická věta

  - Anglická věta z pohledu analytické roviny – každé slovo věty se skládá z několika částí oddělených „|“: slovo ve tvaru vyskytujícím se v textu, lemma, příznak, index slova ve větě, index předchůdce v analytickém stromě, syntaktická funkce. Např.:

    ```
    He|he|PRP|1|2|Sb saved|save|VBD|2|0|Pred my|my|PRP$|3|4|Atr
    ever-lovin|ever-lovin|NN|4|6|Atr ’|’|’’|5|6|AuxG neck|neck|
    NN|6|2|Obj .|.|.|7|0|AuxK
    ```

  - Anglická věta z pohledu tektogramatické rovniny – každé slovo se opět skládá z několika částí oddělelných „|“: t-lemma, funktor, index ve stromě, index předchůdce, typ uzlu, formém a další atributy. Např.:

    ```
    #PersPron|ACT|1|2|complex|n:subj|n.pron.def.pers|sg|-|-|-|-|
    -|-|-|-|-|inan|-|3|-|-|-|-|0|-|- save|PRED|2|0|complex|v:fin|
    v|-|neg0|ant|ind|decl|-|-|-|-|disp0|-|it0|-|-|res0|-|-|1|-|-
    #PersPron|APP|3|4|complex|n:poss|n.pron...
    ```

  - Korespondence mezi anglickou analytickou a tektogramatickou rovinou pro významová slova.
  - Korespondence mezi anglickou analytickou a tektogramatickou rovinou pro pomocná slova.

- Zarovnání mezi českou a anglickou analytickou rovinou
  - Zarovnání nástrojem GIZZA++ „tam“ z češtiny do angličtiny.
  - Zarovnání nástrojem GIZZA++ „zpět“ z češtiny do angličtiny.
  - Zarovnání nástrojem GIZZA++ symetricky z češtiny do angličtiny.
  - Zarovnání nástrojem GIZZA++ symetricky z angličtiny do češtiny.
- Zarovnání mezi českou a anglickou tektogramatickou rovinou
  - T-zarovnání „tam“ z češtiny do angličtiny.
  - T-zarovnání „zpět“ z češtiny do angličtiny.
  - Další t-zarovnání založené na pravidlech spojující zejména generované uzly.

3.3.3 Plaintext formát

Plaintext formát je z těchto formátu nejjednodušší a obsahuje nejméně informací. Každý řádek textu se skládá ze čtyř sloupců oddělených tabulátorem – ID větného páru, skóre filtrování vět, česká věta, anglická věta. Text v tomto formátu je tedy zarovnán jen na věty, nikoliv na slova.

3.4 Popis českých lemmat

Nyní popíšu, v jakém tvaru se v korpusu vyskytují lemmata v českých větách. Obecně má lemma tvar `lemma_:P1_;P2_,P3_(K)`, kde `lemma` je samotné lemma, `P1` je morfosyntaktický příznak, `P2` je sémantický příznak, `P3` je stylový příznak a `K` je komentář (např. vysvětlivka, způsob odvození). Všechny příznaky i komentář jsou nepovinné. Příznaky tvoří velká i malá písmena. Samotná lemma je ve tvaru `VlastníLemma-[0-9]*`, nepovinný řetězec `-[0-9]*` slouží k odlišení různých homonymních tvarů. Lemmata mohou vypadat například takto:

```
Agned_;Y_,t
dementi_,t
FFUK_:B_;K_;u-(Filozof._fakulta_Univerzity_Karlovy
```

Morfosyntaktický příznak popisuje slovní druh nebo případně jeho bližší určení. Rozlišujeme zde příznak pro zkratku (B), nedokonavé sloveso (T) a dokonavé sloveso (W).

Úkolem sémantického příznaku je zařazení do obecné sémantické kategorie. V korpusu se vyskytují příznaky pro příslušníka národa nebo obyvatele území (E), zeměpisný název (G), chemický pojem (H), společnost, organizaci nebo instituci (K), pojmy z oblasti přírodních věd (L), výrobek (R), příjmení (S), lékařské pojmy (U), křestní jméno (Y), slova z oblasti ekonomie a financí (b), výpočetní techniku a elektroniku (c), technologie (g), právo (j), ostatní vlastní jména (m), specifikaci barev (o), pojmy z oblasti politiky, vlády nebo armády (p), kulturu, vzdělávání a umění (u), sport (w), koníčky, volný čas a cestování (y) a ekologii a životní prostředí (z).

Posledním příznakem je příznak stylový, který popisuje stylové zařazení lemmatu. Lemma může být zastaralé (a), expresivní (e), hovorové (h), slangové (l), nářečné (n), knižní (s), cizí slovo (t), vulgánrní (v) nebo se zastaralým pravopisem či pravopisnou chybou (x) [15].

3.5 Popis morfologických značek v české části

Morfologické značky v české části textů určují slovní druhy, jejich poddruhy a další vlastnosti. Značka se skládá z patnácti znaků, jednoho pro každou její část. Pokud je některá z částí značky neuvedena, je místo toho uvedena pomlčka. Značka může být např. ve tvaru:

```
NNFS1-----A----
VB-S---3P-AA---
```

Značky určují slovní druh, slovní poddruh, rod, číslo, pád, přivlastňovací rod, přivlastňovací číslo, osobu, čas, stupeň, negaci, zda se jedná o aktivum či pasivum, variantu, stylový příznak, apod.[16].

3.6 Popis příznaků v anglické části

Příznak v anglické větě se liší. Příznak je tvořen dvěma až čtyřmi znaky a určuje kategorii slovního druhu, např. řadová číslovka, podstatné jméno v množném čísle, druhý stupeň přídavného jména, základní tvar slovesa, člen, atd.[27].

3.7 Popis syntaktických funkcí

Ke každému slovu jak v češtině, tak i v angličtině je také v analytické rovině přiřazena syntaktická funkce. Může se jednat o označení větných členů (podmět, přísudek, přívlastek, ...) nebo pozice v analytickém stromě (kořen, uzel nezávisící na žádném uzlu, hlavní uzel, ...) [14].

3.8 Chyby v korpusu

Korpus Czeng 1.0 není zcela bez chyb. Na některé chyby autoři sami upozorňují, na další jsem narazila při tvorbě slovníků.

3.8.1 Chyby známé autorům korpusu [10]

Chyby vstupních textů

Mezi tyto chyby patří chyby způsobené špatným rozpoznáním znaků (např. při převodu z pdf obsahujícího obrázky stránek knihy), může se jednat např. o špatně spojená nebo

rozdělelná slova, překlepy. Další chybou je výskyt pomlček navíc, např. u slov, která byla v původním textu rozdělená na dva řádky. Také se nepovedlo odstranit veškerá záhlaví a číslování stránek.

Chyby tokenizace

Může se jednat např. o chybu, kdy `Users|%0` mělo být tokenizováno (rozděleno na slova), ale není. Opačným případem je situace, kdy je záporné číslo rozděleno na dvě slova, přestože by nemělo být.

Chyby lemmatizace

V korpusu se vyskytují chybná lemmata českých slov s pomlčkami (např. fyzikálně-chemický).

3.8.2 Chyby nalezené při práci s korpusem

Při práci s korpusem jsem narazila na další chyby. Tyto chyby jsem objevila po srovnání slovníku vytvořeného z tohoto korpusu a existujícího slovníku díky výpisu dvojic, které v původním slovníku nejsou, případně je tam jen české slovo.

Ve vytvořeném slovníku se občas vyskytují dvojice, kde je na obou stranách české slovo (např. akcička, akcička). Vyhledala jsem si výskyt v korpusu a zjistila jsem, že v korpusu se vyskytuje věta: `Vstup 40,- Bezva akcička, na které si v Coca Cola obýváku namícháš nápoj dle vlastní fantazie.` jak v české, tak i v anglické části. Takovýchto chyb jsem objevila více.

Abych zjistila rozsah takových chyb, vyhledala jsem všechna slova s diakritikou v anglické části výše zmíněných výsledků srovnání. Pro tyto dvojice jsem vyhledala věty z paralelního korpusu a také jejich typ zdroje.

Poté jsem nalezené věty procházela a zkoumala, kde se jedná o chybu a kde nikoliv. Zjistila jsem, že některá slova s diakritikou jsou slova francouzská používaná v angličtině. Jinde se jedná o různá vlastní jména, ta se vyskytují hlavně ve větách evropské legislativy týkajících se České republiky (např. České Budějovice). V evropské legislativě se také vyskytují v anglických větách slova, která se do angličtiny nepřekládají (např. štramberské uši, tvarůžky). Dalším případem výskytu diakritiky je chyba v daném slově a diakritika tam evidentně nepatří.

Většina chyb, kdy je v české i anglické části stejná česká věta se vyskytuje u vět pocházejících z paralelních webových stránek. Tyto chyby mohly dle mého názoru vzniknout z důvodu automatického vyhledávání těchto webových stránek a získáváním vět z nich.

Takové české věty na anglické straně jsem nalezla i v datech z technické dokumentace, kde bych to neočekávala. Takové věty však ani nevypadají jako věty z technické dokumentace, jedná se např. o větu `Zezadu se prodral ke stříbru domácí miláček.` nebo `Které děvčata chtěla dostat šaty?`. Z tohoto vyvozuji, že se jedná o další chybu v korpusu – špatné označení zdroje, ze kterého data pocházejí. Tyto chyby však naštěstí na mou práci, generování slovníku, nemají vliv.

Podařilo se mi najít 24 266 vět, které jsou totožné v české i v anglické části. Není však vyloučeno, že se v korpusu takových vět vyskytuje více. Tyto věty pocházejí převážně z paralelních webových stránek (24 231), některé pak z technické dokumentace (35), podle mě však do ní nepatří, jak jsem zmínila výše.

Z důvodu výskytu těchto chyb jsem se rozhodla upravit skripty pro generování slovníků a snažit se chybné věty vynechávat a slova s diakritikou nezahrnovat, více v kapitole týkající se implemetace.

Další typ chyb, na které jsem narazila, je chybné přířazení lemmy. Všimla jsem si, že ke slovu `kolách` ve větě `Přirovnal bych to k tomu, že dříve jezdili lidé do práce na kolách a dnes jezdí do práce opět na kolách.` je přiřazena lemma `cola` namísto správného `kolo` a vzniká tak nesmyslná dvojice. Tato chyba zřejmě vznikla z důvodu nespisovného vyjádření; pokud by se ve větě vyskytoval správný tvar `kolech`, k chybě by nedošlo. Rozsah těchto chyb však nejsem schopna nijak zjistit ani odhadnout.

Kapitola 4

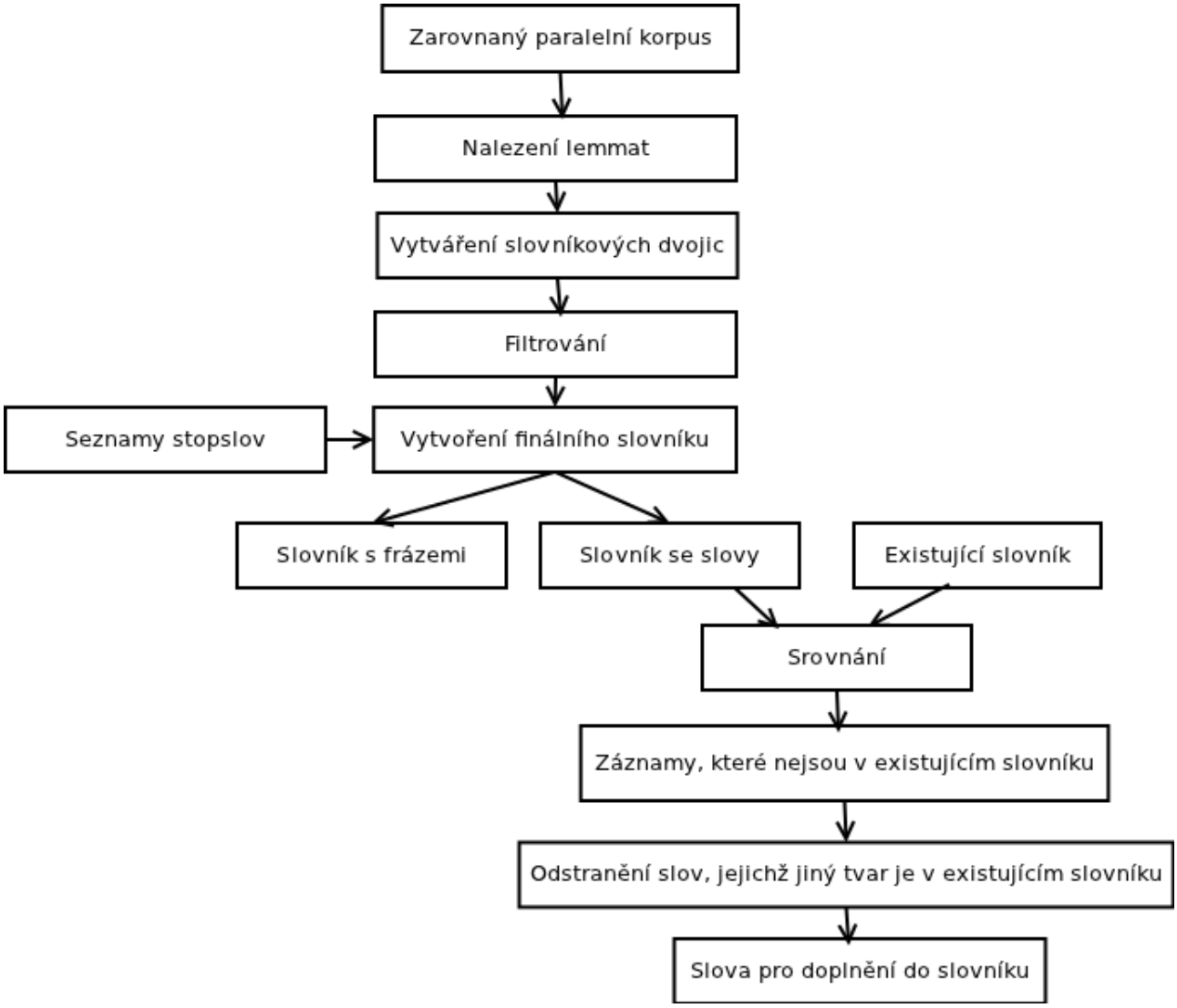
Implementace

Pro implementaci jsem zvolila jazyk Python 3, dále používám shellové skripty pro spouštění skriptů v pythonu a závěrečné úpravy výsledného srovnání slovníků.

Vstupem jsou zarované paralelní texty ve výše popsaném formátu export. Tvorba slovníku se skládá ze dvou fází – výsledkem první fáze je slovník, ve kterém se dvojice slov mohou opakovat a z něj je poté ve druhé fázi vytvořen finální slovník, kde už je každá dvojice jedinečná. Pokud je paralelní text složen z více souborů, první fáze proběhne pro každý soubor zvlášť a poté se výstupy spojí a seřadí do jednoho slovníku.

Výstupem jsou dva slovníky: slovník obsahující pouze samostatná slova jak v české, tak i v anglické části a slovník obsahující vždy buď v české nebo v anglické části víceslovný výraz. Na závěr tyto dva slovníky ještě seřadím podle abecedy a odstraním opakující se záznamy.

Na obrázku 4.1 je schéma systému. Je zpracováván zarovnaný paralelní korpus, v něm jsou nalezena lemmata, vytvořeny dvojice slov a odfiltrována nevhodná slova. Na základě seznamů stopslov se poté vytvoří finální slovník. Vzniknou dva slovníky – slovník se slovy a slovník se slovními spojeními. Slovník se slovy je poté srovnán s existujícím slovníkem; jedním z výstupů je seznam záznamů, které v existujícím slovníku nejsou. Z tohoto seznamu jsou poté odstraněna slova, jejichž jiné tvary v existujícím slovníku jsou, jedná se o slovesa, které mají v původním slovníku variantu se zvratným se a slova, jejichž lemmata nebo odvozené tvary v existujícím slovníku jsou. Takto vznikne další výstup – seznam slov, které je možné doplnit do existujícího slovníku.



Obrázek 4.1: Schéma systému

4.1 Nalezení lemmat

Pro získání lemmatu slouží funkce `findWords()`. Vstupem funkce je slovo z české nebo anglické části věty. Jak jsem popsala výše, toto slovo obsahuje kromě tvaru slova, který se vyskytuje v textu, mimo jiné i jeho lemma, morfologickou značku a index ve větě.

Z tohoto slova získám lemma, které však může obsahovat ještě doplňují informace o slově – morfologický, sémantický a stylový příznak a případně také komentář. Abych dostala samotnou lemmu, použiji regulární výrazy. Výsledkem funkce je tedy lemma v případě, že je vhodná pro zařazení do slovníku, v opačném případě funkce vrací „`0`“. Vybírání vhodných lemmat popisuje další sekce.

4.2 Filtrování

Ne všechna slova jsou vhodná pro zařazení do slovníku. Může se jednat například o vlastní jména, číslice, interpunkční znaménka (v paralelním textu jsou brána jako slova), zkratky. Z tohoto důvodu při extrakci lemmat provádím filtrování slov na základě příznaků, morfologických značek atd.

Nejdříve kontroluji morfologický příznak slova. Podle tohoto příznaku mohu mimo jiné zjistit, že se jedná o zkratku a ty do slovníku nepřidávám.

Poté přichází na řadu sémantický příznak. Slovo vyloučím, pokud se jedná o zeměpisný název, chemický pojem (např. název prvku, sloučeniny), název společnosti, organizace nebo instituce, odborný biologický pojem, název výrobku, křestní jméno nebo příjmení.

Následuje filtrování na základě morfologické značky v případě českého slova a označení slovního druhu v případě slova anglického.

Co se týče českých slov, vynechávám interpunkci, číslovky, částice, citoslovce, slova s neurčeným slovním druhem, dále signaturu autora, slovo „krát“, číslo zapsané číslicemi (jak římskými, tak arabskými).

V případě anglických slov do slovníku nezařazuji číslovky, determinátory (např. the, an, some, this), there vyskytující se v existenční vazbě, cizí slova, značku položky seznamu, symboly, vlastní jména.

Potom provádím kontrolu znaků ve slově – pokud slovo obsahuje nepísmenné znaky, rovněž není do slovníku zařazeno.

Poslední filtrování probíhá na základě délky slov, konkrétně se týká jednopísmenných slov. Pokud slovo obsahuje jen jedno písmeno a nejedná se v případě českého slova o a, i, k, o, s, u, v, z (ať už se jedná o velká či malá písmena [21] nebo v případě anglického slova o a, A, I, O [13], jedná se o neexistující slova, a proto je do slovníku nevkládám.

Jak je popsáno výše, funkce pro získání lemmy v případě nevhodných slov vrací místo lemmy příznak, že slovo bylo odfiltrováno. V případě víceslovných spojení je toto spojení platné jen v případě, že neobsahuje žádné odstraněné slovo.

4.3 Vytváření dvojic na základě zarovnání

Hlavní část skriptu obsahuje spojení slov (slovních spojení) do dvojic na základě informací v zarovnaném paralelním textu a vytvoření slovníku z těchto dvojic. Jak jsem popsala výše, paralelní text se zpracovává postupně po větách.

Poté, co dříve popsaným způsobem získám lemma, uložím jej do pole – pro česká slova `czWords`, pro anglická `engWords`. Index slova v poli je dán jeho indexem ve větě.

Dále potřebuju znát zarovnání – to získám ze třináctého sloupce věty paralelního textu. Tento sloupec obsahuje dvojice indexů slov oddělené mezerami. Tyto indexy získám pomocí funkce `split()` a regulárních výrazů.

Když znám indexy dvojic, které k sobě patří, můžu tato slova přiřadit k sobě. Na základě indexů vyhledám příslušná slova v polích `czWords` a `engWords`. Pro ukládání odpovídajích si slov (slovních spojení) v rámci jedné věty používám pole `czeng`, jehož prvky jsou rovněž pole, dvojice české slovo, anglické slovo.

Pokud by slova byla vždy zarovnaná 1:1, bylo by jejich spojení jednoduché. Často je však na sebe zarovnáno více slov. Tento problém jsem původně řešila tak, že slova spojím „každé s každým“. Tímto sice vznikly správně zarovnané dvojice slov, ale kromě toho také hodně chybných dvojic, kde je například předložka, zvratné se nebo člen k významovému slovu slovního spojení. Proto jsem od tohoto způsobu řešení nakonec upustila. Místo toho jsem se rozhodla spojovat vícenásobně zarovnaná slova do slovních spojení. V konečné fázi potom vzniknou dva slovníky vzlášť pro samostatná slova a zvlášť pro slovní spojení.

4.4 Vytváření slovních spojení

Jak jsem již zmínila v předcházející kapitole, vytvářím nejenom slovník dvojic slov, ale také slovník slovních spojení. Slovní spojení vytvářím spojováním slov, která jsou na sebe zarovnaná.

Pro indikaci, ke kterým českým slovům už je přiřazeno slovo anglické, a naopak, používám dvě pole indexů těchto slov. Pokud při procházení indexů zarovnání narazím na index, který se v alespoň jednom z těchto polí už vyskytuje, znamená to, že dojde k vytvoření slovního spojení na české nebo anglické straně slovníku. Pokud se ani jeden z indexů nevyskytuje v odpovídajícím poli, uložím do slovníku pouze samostatné slovo a indexy pak přidám do polí indexů.

Pokud je českému slovu už přiřazeno slovo anglické, naleznu toto české slovo ve slovníku, zjistím index příslušné dvojice. Poté ke stávajícímu anglickému slovu (slovům) připojím další (pokud v dané dvojici ještě není) a vytvořím tak slovní spojení. Analogicky nakládám s anglickými slovy, kterým už bylo přiřazeno slovo české.

Výše jsem zmínila, že funkce získávání lemmat vrací „`0`“ v případě, že toto slovo není vhodné pro zařazení do slovníku. Protože se poté s touto „`0`“ nakládá stejně jako s platnými slovy, tj. je také přidávána do slovních spojení, je třeba na závěr ještě před přidáním do celkového slovníku tato neplatná slova zohlednit. Původně jsem už při spojování slov do dvojic na základě zarovnání tato slova ignorovala, ale v případě slovních spojení tak vzniklo hodně nesmyslných spojení, protože v nic chyběla slova, proto teď nejdřív vytvořím slovní spojení i s „`0`“ a až poté to řeším. Pokud slovní spojení obsahuje alespoň jednu „`0`“, do celkového slovníku jej nezahrnu.

Když jsem pro vytváření slovních spojení použila stejně jako pro slova lemmata, tato spojení neměla smysl. Proto jsem se rozhodla v případě slovních spojení zachovat původní tvar slov. Jelikož ale v době extrakce lemmatu ještě nevím, zda se bude jednat o samostatné slovo či spojení více slov, musí být možné později použít lemma i tvar použitý v textu. Pro tento účel jsem upravila funkci `findWords()`. Tato funkce nyní nevrací samotné lemma, ale také tvar slova z textu, tyto řetězce jsou odděleny „`|`“. Těsně před přidáním dvojice do slovníku spočítám slova v české i anglické části. Pokud některá z těchto částí obsahuje více slov, použiji v tomto spojení slova ve tvaru, v jakém se vyskytla v textu. V případě samostatných slov volím lemma.

4.5 Porovnávání slovních druhů

I po odfiltrování nevhodných slov obsahoval výsledný slovník hodně chybných a nesmyslných překladů. Z tohoto důvodu jsem se rozhodla filtrovat dvojice i na základě slovních druhů – pokud je každé slovo z dvojice jiného slovního druhu, dvojice se do slovníku nezařadí. Tato kontrola probíhá na závěr zpracování každého řádku před přidáním záznamů do celkového slovníku. Týká se pouze samostatných slov (na české i anglické straně slovníku), víceslovné výrazy se z pohledu slovních druhů nekontrolují.

Abych mohla tuto kontrolu provést, musím nejdřív zjistit slovní druhy slov. Tato informace je obsažena v paralelním korpusu. Co se týče českých slov, je tato informace obsažena v morfologické značce, jak už jsem zmínila dříve. Tato značka obsahuje přímo informaci, o jaký slovní druh se jedná (nebo zda se jedná o interpunkci nebo nerozeznatelný slovní druh). Označení slovních druhů u slov anglické části se liší. Značky neurčují přímo slovní druh, ale podkategorie slovních druhů (například různé časy sloves, rozdělení podstatných jmen na jednotné a množné číslo, atd.). Z těchto podkategorií však rovněž mohu určit slovní druh.

Určování slovních druhů jsem přidala do funkce `findWords()` získávající lemmu. Na základě již zmíněných příznaků se určí slovní druh a jeho pořadové číslo je přidáno do výsledného řetězce, stejně jako tvar slova vyskytující se v textu, jak jsem popsala v předcházející sekci; od lemmatu je toto číslo rovněž oddělelno „`|`“.

Po tomto filtrování však byly odstraněny i správné dvojice. Jednalo se o případy, kdy anglické slovo bylo sloveso ve tvaru gerundia a české slovo podstatné jméno. Druhým případem byla dvojice anlického slovesa v minulém čase a českého přídavného jména. Z tohoto důvodu u anglických slov, pokud se jedná o gerundium nebo sloveso v minulém časem přidávám k označení slovního druhu navíc číslici „`1`“ nebo „`2`“. Pokud při filtrování slovní druhy nesouhlasí, provádím druhé porovnání na základé těchto nově přidaných značek.

4.6 Odfiltrování českých slov v anglické části

Jak jsem již zmínila v předchozí kapitole, není paralelní korpus CzEng zcela bez chyb. Mezi jeho chyby patří výskyt českých vět v anglické části korpusu. Po analýze výsledku srovnání slovníků jsem se rozhodla tento problém řešit, abych zabránila nesmyslným dvojicím.

Potřebovala jsem však najít způsob, jak poznat, že se jedná o českou větu. Vycházela jsem z toho, že česká slova, narozdíl od anglických, mohou obsahovat diakritiku. Proto jsem přidala do funkce `findWords()` kontrolu, jestli dané anglické slovo neobsahuje diakritiku. Pokud ano, do korpusu jej nezařazuji. Tím zajistím, že se v korpusu nebudou vyskytovat slova s diakritikou. Nejedná se sice pouze o česká slova, občas se v anglické větě vyskytne i slovo přejaté z francouštiny, které může obsahovat `é`, ale to není až tak časté. Pokud se jedná o slova s diakritikou, česká slova převládají.

Problém je v tom, že tímto způsobem nevyfiltruji všechna česká slova v anglické části. Česká slova neobsahující diakritiku ve slovníku nadále zůstanou. Napadla mě možnost, že bych v případě výskytu slova s diakritikou do slovníku nezařadila celou větu. Toto řešení mi však nepřišlo zcela vhodné. Jedním z důvodů byl výskyt výše zmíněných slov pocházejících z francouštiny. Dále se mohou v anglických větách, které jsou v pořádku, vyskytovat česká slova, jako jsou místní názvy nebo slova, která se nepřekládají (`tvarůžky`). Tato slova se vyskytují zejména v evropské legislativě. Pokud bych odstranila ze slovníku celou větu, přišla bych i o slova, která jsou v pořádku.

Nakonec jsem tento problém vyřešila tak, že odstraním věty, kde aslespoň čtvrtina slov obsahuje diakritiku. Tuto hranici jsem odhadla na základě několika českých vět. Pokud se bude jednat o ojedinělé české slovo, slovní spojení v anglické větě, toto slovo (slova) se sice vyloučí, zbytek věty zůstane. Je sice možné, že i takto se do slovníku mohou dostat česká slova (je-li v české větě málo slov s diakritikou), nebo se naopak do slovníku některá slova, která by v něm mohla být, nezařadí (příliš mnoho českých slov v jinak anglické větě), přesto mi však tento přístup připadal nejvhodnější. Myslím, že se mi takto povedlo odstranit maximum českých slov z anglické části slovníku a zároveň neodstraňovat zbytečně mnoho správných anglických slov.

Rozpoznání slov s diakritikou jsem ponechala ve funkci `findWords()`. Pouze jsem upravila její návratovou hodnotu v případě nalezení slova s diakritikou. Místo obvyklého „`0`“ vrací v tomto případě „`0d`“. Po návratu z funkce se návratová hodnota testuje a pokud se na tuto hodnotu narazí, nahradí se „`0`“. Kromě toho se však zvýší hodnota proměnné s počtem slov s diakritikou ve větě. Po zpracování celé anglické části se otestuje poměr celkového počtu slov a slov s diakritikou. Pokud diakritiku obsahuje čtvrtina a více z celkového počtu slov, věta se do slovníku nezařadí.

Aby se našla všechna slova s diakritikou, je třeba ve funkci `FindWords()` provést test na diakritiku před dalším filtrováním. V opačném případě by mohlo být slovo s diakritikou vyloučeno z jiného důvodu (např. se může jednat o vlastní jméno) a funkce by vrátila „`0`“. Slovo by se tak nezapočítalo jako slovo s diakritikou a výsledek porovnání s celkovým počtem slov by byl poté zkreslený a nesprávný.

4.7 Vytváření finálního slovníku

Výstupem skriptu `makeDict.py` je, jak už jsem uvedla výše, slovník s opakujícími se dvojcemi. Tento slovník je vstupem pro druhý skript, `uniqDic.py`, který vytváří finální slovník, v němž jsou již dvojice unikátní. Jeho výstupem jsou dva slovníky, zvlášť slovník pro slovní spojení a pro slova.

Součástí tohoto skriptu je i počítání výskytů dvojic. Pokud je spuštěn s přepínačem `-c`, vypíše na výstup kromě dvojic slov i počet výskytů této dvojice. Tato funkcionalita byla do skriptu přidána za účelem analýzy výsledků.

Vstupní soubor procházím po řádcích. Z důvodu zmíněného počítání výskytů používám pro každé nové české slovo (slovní spojení) proměnnou `czeng`, která má datový typ slovník. Jedná se o uspořádanou kolekci dvojic klíč-hodnota chovající se jako vyhledávací tabulka. V tomto případě je klíčem anglické slovo a hodnotou počet výskytů. Pokud už anglické slovo ve slovníku je, počet výskytů pro toto slovo se zvýší o jedna.

V okamžiku, kdy je na vstupu dvojice s jiným českým slovem, vkládám záznamy do finálního slovníku. V průběhu tohoto vkládání probíhá poslední filtrování záznamů, tentokrát na základě stopslov. Poté se na základě počtu slov rozhodne, do kterého ze slovníků se daná dvojice přidá, případně se ke dvojici vypíše i počet výskytů.

Poté se vše opakuje pro další české slovo ze vstupního slovníku.

4.8 Stopslova

Abych eliminovala co nejvíce chybných záznamů nehodících se do slovníku, rozhodla jsem na závěr tyto záznamy filtrovat na základě seznamů stopslov. Stopslovo je slovo, které se v daném jazyce často vyskytuje, ale nenese významovou informaci. Stopslova mají často jen

syntaktický význam. Může se jednat o předložky, spojky, zájmena, malovýznamová slovesa, členy atd. [23]. Při vytváření seznamů stopslov jsem se inspirovala seznamy nalezenými na internetu [17], [8] a [1], seznamy jsem dále upravila.

Dalším vstupem druhého skriptu jsou tedy dva soubory se stopslovy – českými a anglickými. Seznam stopslov není tedy pevně daný v programu. Na začátku skriptu `uniqDict.py` načtu tato stopslova a uložím je do polí `stop_cz` a `stop_en`.

K filtrování na základě stopslov dochází v okamžiku vkládání záznamů do finálního slovníku (výpisu na výstup skriptu `uniqDict.py`). Než se daná dvojice vypíše, ověří se, jestli je vhodná pro zařazení do slovníku. Pokud se jedná o jedno slovo, je vhodné pro zápis, pokud to není stopslovo. Dále také do slovníku nezařazuji slovní spojení, která se skládají pouze ze stopslov. Dvojici do slovníku zařazuji v případě, že do něj můžu zařadit jak anglické, tak české slovo (slovní spojení).

Dále před vkládáním do slovníku odstraňuji zbytečná nevýznamová slova na začátku slovních spojení. Toto odstraňování probíhá na základě dvojic stopslov uložených v polích `cz2` a `en2`. Slova do těchto seznamů jsem se snažila volit tak, abych z co nejvíce nevhodných záznamů vytvořila záznamy vhodné. Pokud pak slovní spojení v jednom jazyce začíná na slovo ze seznamu pro tento jazyk (např. a, ale, já), ale jeho ekvivalent ve druhém jazyce na odpovídající slovo (např. and, but, I) nezačíná, z prvního slovního spojení toto slovo odstraním. Takto se z frází zbavím zbytečných slov, která nenesou žádný význam, a zároveň však tato slova zachovám v případě, že se vyskytují v obou dvojicích a mohlo by se jednat o ustálenou frázi.

Kapitola 5

Popis skriptů

5.1 Tvorba slovníku

5.1.1 Vstup

Nejdůležitějším vstupem systému pro tvorbu slovníků jsou samozřejmě paralelní texty. Tyto paralelní texty musí být ve výše popsaném export formátu a musí se nacházet všechny v jednom adresáři. Kromě toho jsou vstupem ještě dva soubory – soubor s českými stopslovy a soubor s anglickými stopslovy. Každý řádek těchto souborů je stopslovo v příslušném jazyce.

5.1.2 Výstup

Výstupem jsou dva soubory se slovníkem. První z nich je slovník, kde anglická i česká část obsahuje vždy jen samostatné slovo, druhý je slovník, kde je alespoň jedna část dvojice tvořena více slovy.

5.1.3 Soubory

Systém tvoří soubory `makeDict.sh`, `makeDict.py` a `uniqDict.py`.

Soubor `makeDict.py` je skript v jazyce Python 3. Tento skript vytváří první verzi slovníku, mezikrok, kdy se ve slovníku dvojice opakují. Vstupem skriptu je soubor se zarovnaným paralelním textem v export formátu, výstupem výše popsaný slovník.

Dalším souborem je `uniqDict.py`. Opět se jedná o skript v jazyce Python 3. Úkolem tohoto skriptu je vytvořit konečný slovník, kde se dvojice neopakují. Na vstupu skript očekává slovník s opakujícími se dvojicemi a dva soubory se stopslovy, výstupem jsou dva slovníky – slovník se slovy a slovník s frázemi. Pokud je skript spuštěn s přepínačem `-c`, vypisuje ke dvojicím počet jejich výskytů.

Soubor `makeDict.sh` je shellový skript, který celé generování slovníků řídí, spouští předchozí dva skripty. Na vstupu očekává adresář s paralelními texty a soubory se stopslovy, výstupem jsou dva slovníky seřazené podle abecedy. Pro každý soubor s paralelním textem ze zadaného adresáře je spuštěn skript `makeDict.py`. Výsledné slovníky jsou spojeny do jednoho souboru, ten je seřazen a je vstupem pro následovně spuštěný skript `uniqDict.sh`. Výsledkem jsou dva slovníky, ty jsou poté ještě seřazeny.

5.1.4 Spuštění

Spuštění skriptu `makeDict.sh` je ve tvaru:

```
./makeDict.sh data stop_cz stop_en dict_words dict_phrases
```

kde jednotlivé parametry znamenají:

- `data` – adresář s paralelními texty
- `stop_cz` – soubor s českými stopslovy
- `stop_en` – soubor s anglickými stopslovy
- `dict_words` – výstupní soubor – slovník se slovy
- `dict_phrases` – výstupní soubor – slovník se slovními spojeními

Kapitola 6

Výsledek

Výstupy skriptu jsou dva slovníky - slovník obsahující na obou stranách pouze slova a slovník s víceslovnými výrazy. Slovník se slovními spojeními obsahuje 4 169 780 záznamů. Je v něm však hodně nesmyslných dvojic, vzniklých kvůli zarovnání `m:n`. Často se stává, že záznamy obsahují slova navíc. Když slovník s frázemi seřadím podle četnosti výskytu, nejpočetněji zastoupené fráze jsou většinou správné, zatímco fráze s nízkou četností správné nejsou. Proto by podle mého názoru bylo možné vybrat ze slovníku jen některé fráze, určit nějakou mezní hodnotu četností a fráze s četností nižší do slovníku nezařazovat.

Správně přeložené fráze jsou např.: `zejména - in particular`, `podle - according to`, `alespoň - at least`, `lze - may be`, `například - for example`, `dnes večer - tonight`. Jako příklad frází, kde je (jsou) slovo (slova) navíc (nejčastěji se jedná o předložky), mohu uvést: `kvalita - quality of`, `vydaný - issued by`, `číslo - document number`, `pošta - by post`, `brzy - would soon`, `zapomněl úplně - forget`. Jak už jsem zmínila, slovník obsahuje i úplně nesmyslné záznamy, jako např.: `udržet - check her`, `jít - 's fun`, `bylo ticho - pass`, `a abych - and getting`, `aba - backwardness to`, `gladiátor čísi - rear`.

Slovník se samostatnými slovy tolik nesmyslných záznamů neobsahuje. Proto jsem se více věnovala zhodnocení tohoto slovníku a jeho srovnání s existujícím slovníkem. Tento slovník obsahuje 920 184 slov. Také tento slovník obsahuje i nějaké chybné překlady, např: `zařízení - art`, `věnovat - push`, `šetřit - know`, `relaxace - body`, `poštvat - play`. Přesto obsahuje mnohem více správných dvojic, jak vyplynulo ze srovnání s existujícím slovníkem.

Pro srovnání jsem využila slovníky uložené na školním serveru minerva v adresáři `/mnt/minerva1/nlp/projects/dicts2lmf`, konkrétněji se jednalo o jeden anglicko-český slovník a dva česko-anglické. Tyto slovníky jsem nejdřív převedla z formátu lmf (formát založený na xml) do zjednodušeného tvaru (na každém řádku jeden záznam, odděleno tabulátorem) a spojila do jednoho česko-anglického slovníku, aby bylo srovnání jednodušší. V tomto vzorovém slovníku je 926 367 záznamů.

Výsledkem srovnání bylo následujících 5 souborů (jsou uloženy v adresáři /mnt/minerva1/nlp/projects/mt_dict/results):

- `both` – dvojice, které se celé vyskytují v obou slovnících
- `cz_en_diff` – dvojice z nového slovníku, kde se obě slova v původním slovníku vyskytují, ale každé v jiné dvojici
- `cz_only` – dvojice z nového slovníku, kde se v původním vyskytuje jen české slovo

- `en_only` – dvojice z nového slovníku, kde se v původním vyskytuje jen anglické slovo
- `nothing` – dvojice, které se v původním slovníku nevyskytují

Bylo nalezeno 144 004 záznamů obsažených v obou slovnících. Z 656 139 záznamů ve vytvořeném slovníku jsou v existujícím slovníku obě slova, každé se ale vyskytuje v jiné dvojici. Z 67 286 záznamů ve vytvořeném slovníku je v původním jen české slovo, z 46 561 jen anglické, 6 158 záznamů se v původním slovníku vůbec nevyskytuje.

Ze souborů s výsledkem srovnání jsem poté vytvořila jejich upravené varianty, kde jsou dvojice slov seřazené podle četnosti výskytu, která je v těchto souborech obsažena. Kromě toho je uveden i počet dokumentů výskytu a jeden příklad věty v obou jazycích ze vstupního paralelního korpusu. Je tedy možné vytvořit slovník obsahující příklady použití v něm obsažených slov. Příklad takového záznamu:

```
zobrazitelný displayable 4 4 Zprávu nebylo možno reprodukovat
v zobrazitelném formátu . Could not render message in a displayable
format .
```

Dále jsem se zaměřila na soubor `nothing`, který obsahuje záznamy, jež se v existujícím slovníku nevyskytují, a lze jej tedy použít pro doplnění záznamů do slovníku. Bylo jej však třeba nejdřív ještě upravit, jelikož obsahoval slova, jejichž jiné tvary, případně lemmata, se ve slovníku vyskytují. K těmto úpravám jsem použila morfologický analyzátor ma pro češtinu a volně dostupný lemmatizátor lemmagen pro angličtinu. Nejdřív jsem odstranila dvojice, které obsahují české sloveso používající se se zvratným se nebo si, pokud je toto sloveso i se zvratným zájmenem ve slovníku obsaženo. Potom jsem odstranila slova, jejichž jiný tvar, popř. lemma, je ve zdrojovém slovníku. Takto jsem získala seznam záznamů, z nichž se ani jedno slovo (a žádný jeho jiný tvar) v původním slovníku nevyskytuje. Těchto záznamů je 2 590.

Na první pohled je patrné, že vytvořený slovník, na rozdíl od referenčního, obsahuje mnoho odborných termínů. Po analýze prvních sto dvojic po seřazení podle počtu výskytů (36 a více výskytů) jsem zjistila, že z těchto 100 pojmů je 70 odborných termínů, z toho 26 je z oblasti medicíny, 26 biologických, 8 informatických a 7 chemických. Z biologických termínů je obsaženo hlavně hodně latinských názvů rostlin nebo živočichů. 9 záznamů jsou názvy nebo vlastní jména, 2 cizí slova na české straně a ve 3 případech se jedná o chybnou lematizaci.

Naopak 71 551 dvojic slov z původního slovníku se ve vytvořeném nevyskytuje. Toto velké číslo je dáno i tím, že vzorový slovník obsahuje i vlastní jména a názvy, které jsem do vytvářeného slovníku nezahrnovala. Dále se často jedná o slova hodně neobvyklá a nepoužívaná (žiratář, zkuřka, skampolo, abaton), z tohoto důvodu se zřejmě v paralelním korpusu nevyskytují.

Kapitola 7

Závěr

Zadáním této bakalářské práce bylo vytvořit systém pro automatickou tvorbu slovníků z paralelních textů. K tomuto účelu byl použit zarovaný paralelní korpus CzEng. Kromě toho jsou v práci popsány související teoretické pojmy z oblasti strojového překladu a je analyzován použitý paralení korpus.

Výstupem práce jsou dva česko-anglické slovníky, slovník obsahující jen samostatná slova (4 169 780 záznamů) a slovník s víceslovnými frázemi (920 184 záznamů). Slovník se slovy jsem dále srovnávala s existujícím referenčním slovníkem, 144 004 záznamů se vyskytuje jak ve vytvořeném, tak v původním slovníku. Dalším výstupem je seznam záznamů pro doplnění do existujícího slovníku. Tento seznam by však před doplněním do slovníku bylo vhodné ještě projít, jelikož obsahuje hodně odborných termínů, zejména z oblasti biologie a medicíny.

Systém by bylo možné ještě vylepšovat. Největší rezervy má ve tvorbě slovníku s víceslovnými výrazy, další práce by se mohly zaměřit na jejich hodnocení a vybírání správných frází, odfiltrování těch nesmyslných. Další možností by bylo rozšiřit systém tak, aby místo práce se zarovnanými texty tyto texty sám zarovnával.

Literatura

[1] Armand Brahaj: List of English Stop Words. `http://norm.al/2009/04/14/list-of-english-stop-words/`, 2009-04-14 [cit. 2013-04-08].

[2] Bojar, O.; Žabokrtský, Z.; Dušek, O.; aj.: The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC2012*, ELRA, Istanbul, Turkey: European Language Resources Association, Květen 2012, in print.

[3] Clark, A.; Fox, C.; Lappin, S.: *The handbook of computational linguistics and natural language processing*, ročník 57. Wiley-Blackwell, 2010.

[4] Hajič, J.; Hajičová, E.; Panevová, J.; aj.: Anglická tektogramatická rovina v kostce. `http://ufal.mff.cuni.cz/pcedt2.0/cs/credits.html`, 2012 [cit. 2013-04-16].

[5] Hutchins, W. J.: Machine translation: A brief history. *Concise history of the language sciences: from the Sumerians to the cognitivists*, 1995: s. 431–445.

[6] Karásek, J.; Šanda, P.; Burget, R.; aj.: Strojové učení základem pro hybridní lemmatizační algoritmus. *Elektrorevue*, ročník 57, 2012, ISSN 1231539.

[7] Kumano, A.; Hirakawa, H.: Building an MT dictionary from parallel texts based on linguistic and statistical information. In *Proceedings of the 15th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, 1994, s. 76–81.

[8] Michal Janík: Stop slova. `http://www.michaljanik.cz/oblibene/stop-slova`, 2009-06-03 [cit. 2013-04-08].

[9] Nakazawa, T.; Kurohashi, S.: Statistical phrase alignment model using dependency relation probability. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*, Association for Computational Linguistics, 2009, s. 10–18.

[10] Ondřej Bojar: CzEng 1.0 Known Issues. `http://ufal.mff.cuni.cz/czeng/czeng10/known-issues.html`, 2011-12-14 [cit. 2013-04-16].

[11] Santos, A.: A survey on parallel corpora alignment.

[12] Véronis, J.: *Parallel Text Processing: Alignment and use of translation corpora*, ročník 13. Kluwer Academic Pub, 2000.

[13] WWW stránky: Category talk:English one letter words.
http://en.wiktionary.org/wiki/Category_talk:English_one_letter_words,
2006-12-26 [cit. 2013-04-08].

[14] WWW stránky: Průvodce českým akademickým korpusem 2.0 : Popis analytických funkcí.
http://ufal.mff.cuni.cz/rest/CAC/doc-cac20/cac-guide/cz/html/ch14.html,
2008 [cit. 2013-04-16].

[15] WWW stránky: Průvodce českým akademickým korpusem 2.0 : Popis lemmat.
http://ufal.mff.cuni.cz/rest/CAC/doc-cac20/cac-guide/cz/html/ch12.html,
2008 [cit. 2013-04-16].

[16] WWW stránky: Průvodce českým akademickým korpusem 2.0 : Popis morfologických značek.
http://ufal.mff.cuni.cz/rest/CAC/doc-cac20/cac-guide/cz/html/ch13.html,
2008 [cit. 2013-04-16].

[17] WWW stránky: Česká stopslova – Webtrh.
http://webtrh.cz/30601-ceska-stop-slova, 2009-03-12 [cit. 2013-04-08].

[18] WWW stránky: Rule based machine translation.
http://language.worldofcomputing.net/machine-translation/rule-based-machine-translation.html, 2009-12-13 [cit. 2013-05-03].

[19] WWW stránky: Example based machine translation.
http://language.worldofcomputing.net/machine-translation/example-based-machine-translation.html, 2010-09-27 [cit. 2013-05-03].

[20] WWW stránky: Encyklopedie Navajo.
http://cs.wikipedia.org/wiki/Encyklopedie_Navajo, 2011-10-28 [cit. 2013-04-16].

[21] WWW stránky: Délka slov v češtině.
http://cs.wikipedia.org/wiki/D%C3%A9lka_slov_v_%C4%8De%C5%A1tin%C4%9B,
2012-02-07 [cit. 2013-04-08].

[22] WWW stránky: Opus, the open parallel corpus. http://opus.lingfil.uu.se/,
2013-02-26 [cit. 2013-05-04].

[23] WWW stránky: Stopslovo. http://cs.wikipedia.org/wiki/Stopslovo, 2013-03-09 [cit. 2013-04-08].

[24] WWW stránky: Korpus InterCorp.
http://www.korpus.cz/intercorp/?req=page:info, 2013-04-10 [cit. 2013-05-04].

[25] WWW stránky: JRC-Acquis. http://ipsc.jrc.ec.europa.eu/index.php?id=198,
2013-05-03 [cit. 2013-05-04].

[26] WWW stránky: hunalign – sentence aligner.
http://mokk.bme.hu/resources/hunalign/, 2013 [cit. 2013-05-04].

[27] WWW stránky: Penn Part of Speech Tags.
http://cs.nyu.edu/grishman/jet/guide/PennPOS.html, [cit. 2013-04-16].

[28] WWW stránky: Průvodce PDT 2.0 : rovniny anotace.
http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/cz/html/ch02.html, [cit. 2013-04-16].

[29] WWW stránky: European Parliament Proceedings Parallel Corpus 1996-2011s.
http://www.statmt.org/europarl/, [cit. 2013-05-04].